DOCUMENT RESUME

ED 164 603 TM 008 167

AUTHOR Forster, Fred; And Others
TITLE Research on the Rasch Measurement Model.
PUB DATE Mar 78
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; Complexity Level; Elementary Secondary Education; *Equated Scores; *Item Analysis; *Item Banks; Mathematical Models; Research Reviews (Publications); *Sampling; Test Items
IDENTIFIERS *Rasch Model; *Rasch Scaled Scores

ABSTRACT

Research on the Rasch model of test and item analysis was applied to tests constructed from item banks for reading and mathematics with respect to five practical problems for scaling items and equating test forms. The questions were: (1) Does the Rasch model yield the same scale value regardless of the student sample? (2) How many students are needed to accurately scale a test: (3) Are the scale values for items equal interval? (4) Does the Rasch produce equal achievement levels for the same students on different tests? (5) Does an item receive the same scaling regardless of the other items on the test? All of the research, including replications of three of the studies, supported the usefulness and accuracy of the Rasch model. It was judged that 200 students were sufficient to produce reliable item scaling. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PRINTED IN THE U.S.A.

Research on the Rasch Measurement Model

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED164603

Fred Forster
Portland Public Schools

Gordon Ascher
State Department of Education

Chadwick Karr
Portland State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Fred Forster

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Abstract

Over the past four years, the Northwest Evaluation Association has worked toward the development of test item banks based on the Rasch measurement model. Among the research questions which have been explored in this effort are the following:

(1) Are Rasch item difficulties the same when calculated for different groups of students?

(2) Do Rasch item difficulty levels form an equal interval scale?

(3) Do students receive the same Rasch achievement level when they are measured by different tests?

The paper presents empirical evidence concerning these and other research questions and presents the practical importance of this research for designing school testing programs.

IM008 167

PRINTED IN THE U.S.A.

Research on the Rasch Measurement Model

Fred Forster
Portland Public Schools

Gordon Ascher
State Department of Education

Chadwick Karr
Portland State University

Background

It is twenty years since Georg Rasch first proposed his measurement model and fifteen since Benjamin Wright began his almost single-handed fight to bring the benefits of this model to the fore. During this time the model has been steadfastly ignored by nearly all prominent measurement authorities and test publishers in spite of its great promise. Even more unfortunately, the literature on the Rasch is dominated by studies on computer generated data and exotic applications with only a handful of studies drawn from actual school settings.

Five years ago, our district, together with fifty others in the states of Washington and Oregon, began working on the development of item banks in several school subjects, beginning with reading, mathematics and language arts. This consortium, the Northwest Evaluation Association, immediately faced the problems of coming up with a flexible method for scaling items and equating test forms which did not require the costs involved in the extensive testing of random samples. When we found that the Rasch technique promised to do just that, we felt it was a natural solution to our problems. Unfortunately, we also found that virtually no practical research had been done to help us start our work on the reading and mathematics banks. This need launched our research.

Overview

The research described here was aimed at the following practical questions:

(1) Does the Rasch model yield the same scale values regardless of the student sample?
(2) How many students are needed to accurately scale a test?
(3) Are the scale values for items equal interval?
(4) Does the Rasch produce equal achievement levels for the same students on different tests?
(5) Does an item receive the same scaling regardless of the other items on the test?

We believed that each of these questions was extremely important in deciding whether the Rasch would be of real value in our work.

---

*Presented at the A.E.R.A. Annual Conference, Toronto, March 1978

Research

This section will present each research questions and then indicate the evidence we gathered to test it.

Question: Does the Rasch model yield the same scale value regardless of the student sample?

Method: 1500 students at grade four took the district standardized achievement test in reading in the Fall of 1973.
The students were subdivided into the following groupings: (1) female students scoring below the district average, (2) male students scoring below the average, (3) female students scoring above the average, (4) male students scoring above the average, (5) the 15 highest scoring schools, and (6) the 15 lowest scoring schools. The items on the tests were scaled separately for each of these groupings as well as for the total group of 1500 students. The resulting item scale values were correlated to determine whether the different student samples yielded the same scaling.

Results: As shown in Figure 1, the scaling for each subgroup agreed with that for the total group.

Replication: The study was replicated using data from 4500 students who took the district reading and mathematics tests at grades 4 and 8 in the Fall of 1976. As shown in Figure 2, the results supported those from the original study.

Question: How many students are needed to accurately scale a test?

Method: 1500 students in grades four and eight took the district standardized tests in reading and mathematics in the Fall of 1976. Five random samples of each of the following approximate sizes were drawn: 300, 200, 100, and 50. Each of the random samples was scaled independently and then correlated with the scaling for the total group.

Results: As shown in Figure 3, the correlations between the scaling for the samples and for the total group were quite high. On the other hand, when the ratio of the standard deviation of the item scaling for each sample was compared to that for the total group to determine the degree to which the scalings were of equal magnitude, there were differences. As shown in Figure 3, the samples of 200 and 300 more closely reflected the calibrations for the total group than the smaller samples. In consequence, a sample size of 200 appears to be sufficient to produce reliable item scaling.

Replication: This study was replicated using 3800 students at grades four and seven who took the district reading and mathematics standardized tests the Fall of 1976. As shown in Figure 4, this study also supported the 200 figure.

Question: Are the scale values for items equal interval?

Method: Seven mathematics tests at the upper grade level were constructed

from 250 previous field tested items, and administered to students in grades six through nine. The tests were tied together by including common items within forms to provide two independent estimates of the consistency of the scaling of the tests.

Results: As shown in Figure 5, the scalings from the two independent linking paths yielded comparable results indicating that the scaling of the tests was consistent.

Question: Does the Rasch produce equal achievement levels for the same students on different tests?

Method: Approximately 200 students in grades four through eight were given a short field test along with their standardized reading achievement test in the Fall of 1976 (approximately 200 students for each ten pairings of tests). The students were scored separately on each test, and the resulting achievement levels were compared.

Results: As shown in Figure 6, the agreement between the student's Rasch achievement levels on the two tests was high, in some cases exceeding that between the standardized test and its own vocabulary and comprehension subcomponents. These results support the contention that students will receive equal Rasch achievement levels on different tests.

Replication: In the Fall of 1978, approximately 4500 students at each grade level, grades four through eight, took two forty-item tests in reading and two in mathematics. They were scored separately on each test and these scores were converted into Rasch achievement levels. As shown in Figure 7, the results from this study indicated that students received comparable achievement levels from two tests when the tests were both appropriate. At the extremes, the CP tests did not not match the students, and therefore, did not provide the same quality of measurement.

Question: Does an item receive the same scaling regardless of the other items on the test?

Method: In Fall 1975, 1500 students took the districts fourth grade mathematics test which was eighty items long. In addition to scaling all the items, scaling was run eliminating the ten highest and lowest level items, i.e., for the sixty middle level items. Subsequent scalings were made for the middle forty, twenty, and ten items. The item scalings for each of these subscalings was then compared to that for the full eighty items.

Results: As shown in Figure 8, the correlations, even when restricted to force the metrics to be identically equal, were all close to 1.00 indicating that the decreasing range of item levels had negligible effect on the item scalings.

Practical Implications

This research has had a direct impact on the development of tests in our district. The findings for Question 3 led us to design our field tests according to a pattern of overlapping items which makes it possible

to cross validate every phase of our item scaling. Based on the findings for Question 2, we used at least 200 students to field test a test form, drawing students from those classrooms where teachers and principals express a desire to participate in the field testing. In developing our district testing program, we have taken advantage of the research for Question 4 to tie to past data for our district even though we completely renovated the program.

If you are interested in a more thorough discussion of the research presented in this paper as well as related developments, a series of monographs is available though the Northwest Evaluation Association at a cost of ten dollars, by writing to:

Dr. Marlyn Williardson
Clackamas County Intermediate Education District
Marion Hall, Marylhurst Campus
Marylhurst, Oregon 97036

Figure 1

Comparison of Rasch Item Levels Calibrated on Student Subgroups

COMPARISON	CORRELATION
Girls above the average vs. Total	.973
Girls below the average vs. Total	..958
Boys above the average vs. Total	.969
Boys below the average vs. Total	.952
Inner-City Schools vs. Total	.986
Top Schools vs. Total	.983
Inner-City Schools vs. Top Schools	.965

Figure 2

Correlations of Item Levels for Subgroups and the Total Group

	SUBJECT			
	4th Reading	4th Math	8th Reading	8th Math
Group	r	r	r	r
Above The Average vs. Total	.989	.981	.983	.970
Below The Average vs. Total	.964	.974	.976	.953
Inner-City Schools vs. Total	.998	.994	.996	.995
Top Schools vs. Total	.998	.990	.998	.996
Inner-City Schools vs. Top Schools	.993	.984	.990	.983

Figure 3

Item Levels as Calibrated for Different Student Samples

	FOURTH MATHEMATICS			EIGHTH READING		
	80 items 1478 Students σ(Item level)= 14.15			100 Items 1808 Students σ(Item Level) = 15.25		
Sample Size*	Correlation	Ratio of Standard Deviation	Average Discrepancy	Correlation	Ratio of Standard Deviation	Average Discrepancy
50	.956	1.102	3.1076	.965	1.034	3.042
100	.978	1.029	1.9492	.982	1.019	2.074
200	.987	1.107	1.3138	.989	.994	1.381
300	.989	1.010	1.1378	.991	.996	1.103

*Entries represent the average for five samples drawn of each size. Relationships are derived for the item levels calculated for each sample.

Figure 4

Item Levels as Calibrated for Different Student Samples

	FOURTH READING			FOURTH MATHEMATICS			SEVENTH READING			SEVENTH MATHEMATICS		
Number of Items	35			37			36			30		
Total Population	3771			3827			3846			3773		
Standard Deviation	13.51			11.84			6.17			8.30		
Sample Size*	Corr.	Ratio	Avg. Disc.	Corr.	Ratio	Avg. Disc.	Corr.	Ratio	Avg. Disc.	Corr.	Ratio	Avg. Disc.
50	.956	1.045	3.367	.950	1.100	3.291	.939	1.099	2.689	.901	1.205	2.531
100	.977	1.019	2.430	.975	1.024	2.201	.961	1.104	2.086	.941	1.096	1.884
150	.986	.990	1.822	.985	1.007	1.577	.978	1.020	1.463	.955	1.087	1.639
200	.990	.992	1.623	.990	1.021	1.388	.983	1.014	1.299	.971	1.038	1.244
250	.992	1.006	1.353	.992	1.020	1.173	.982	1.054	1.407	.972	1.008	1.219
300	.994	.993	1.148	.994	1.022	1.113	.986	1.045	1.170	.978	.999	1.067

*Entries represent the average for five samples drawn of each size.

Figure 5

Comparison of Item Scaling Across Seven Tests

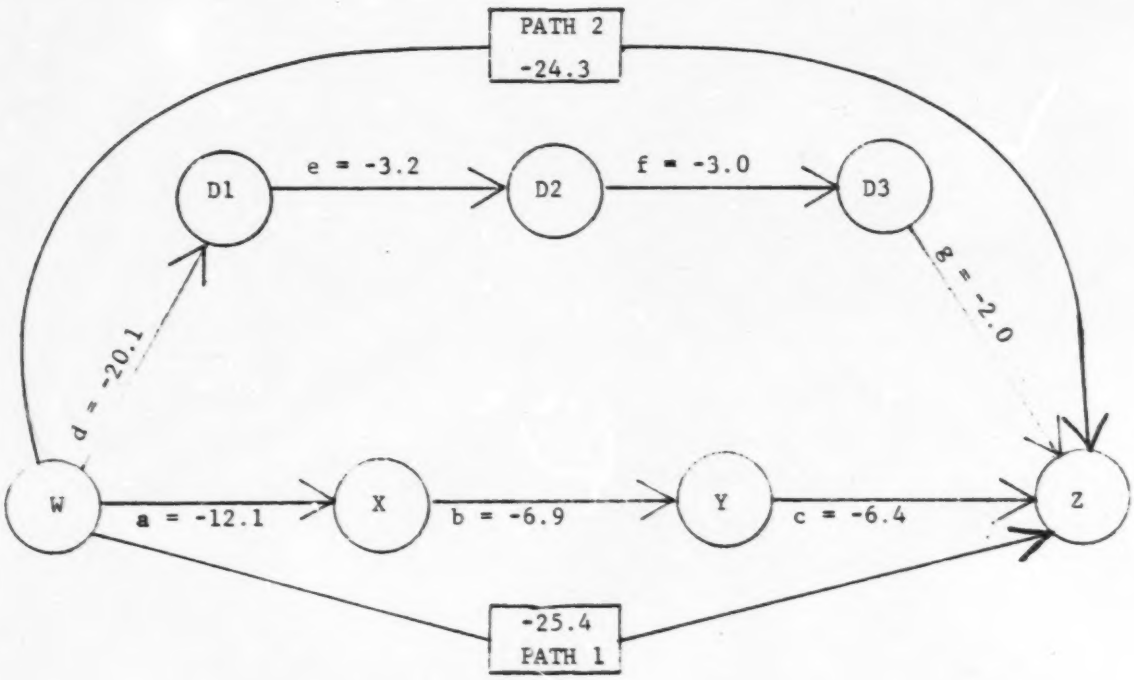


Figure 6

Comparability of Rasch Achievement Levels for a Short Field Test and a Standardized Reading Test

Grade	Field Test Form	Average Raw Score		Average Rasch Achievement Levels	
		Field Test	Standardized Test	Field Test	Standardized Test
3	R12	18.1	16.2	186.8	187.2
4	R17	19.6	33.5	194.9	194.6
4	R35	19.6	36.1	195.8	196.8
5	R36	15.6	31.7	199.3	199.1
6	R25	14.4	38.6	204.4	205.3
6	R37	11.0	38.4	206.2	205.3
7	R27	14.1	41.0	213.1	211.4
7	R38	12.9	35.7	206.6	207.4
8	R32	12.7	61.8	211.7	211.7
8	R39	15.0	57.5	211.2	209.4

Figure 7

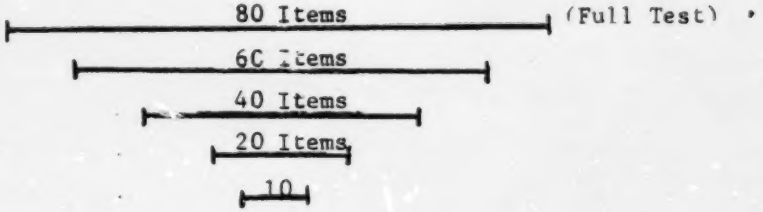
COMPARISON OF RESULTS BETWEEN THE COMPETENCY PROGRESS AND ACHIEVEMENT LEVEL TEST PROGRAMS

READING			MATHEMATICS		
GRADE	CP* AVERAGE	AL** AVERAGE	GRADE	CP AVERAGE	AL AVERAGE
4	194.12	193.76	4	191.28	190.36
5	200.00	200.00	5	200.81	200.00
6	205.53	205.74	6	207.22	208.36
7	209.77	211.21	7	213.49	214.87
8	214.74	216.83	8	222.15	221.41

*Competency Progress
**Achievement Level

Figure 8

Comparison of Item Levels for Subtests of Varying Ranges and Difficulty Drawn from a Fourth Grade Mathematics Test



Subtest	Full Test	60	50	40	30	20	10
60	.998						
50	.996	.999+					
40	.995	.999	.999+				
30	.993	.999	.999+	.999+			
20	.989	.997	.999	.999	.999+		
10	.991	.998	.999	.999+	.999+	.999+	

*Restricted to require equal metrics